3A

УДК 004.89

А.В. Анисимов, К.С. Лиман, А.А. Марченко
Киевский национальный университет им. Т. Шевченко, г. Киев, Украина
lymonadd@gmail.com

Методы вычисления мер семантической близости слов естественного языка

В данной статье приводятся экспериментальные данные вычисления мер семантического сходства и связанности. Все меры, представленные в статье, используют в качестве базы знаний только WordNet. Также авторами были предложены и проверены в эксперименте модификации существующих мер.

Введение

Устранение семантической неоднозначности – это процесс выбора определенного смысла слов исходя из их контекста. В этой задаче важным моментом является определение связанности разных смыслов, поскольку, хотя смысл слова выбирается из некоторого определенного множества, избранное значение слова должно наиболее соответствовать (в семантическом смысле) соседям по тексту, быть связанным с ними, быть семантически похожим. Для этого вводятся различные меры сходства и связанности, которые используются также в таких задачах, как: определение структуры текста, аннотирование и реферирование текстов, информационный поиск, автоматическое индексирование и автоматическая коррекция ошибок в текстах. В данной статье эти понятия различаются на основе [1] следующим образом: сходство – более узкое понятие, похожие сущности обычно связаны одинаковостью по определенной характеристике, а непохожие сущности могут быть семантически связаны другим способом (например: *машина-колесо).*

Для вычисления мер семантического сходства и связанности был разработан программный пакет, который основан на использовании сетевых баз знаний. В данной реализации было использовано лексико-семантическую базу знаний WordNet.

Далее в этой статье будет дано краткое описание структуры WordNet, классическое описание и описание модификаций реализованных мер, таких как: [2-5], и простая мера, пропорционально обратная кратчайшему пути. Затем приведены описание эксперимента и результаты. Эксперимент был поставлен на двух множествах данных – пары английских слов, которым в соответствие вручную были проставлены значения их семантического сходства и связанности. Первая – это множество из 353 пар английских слов, а вторая – 30 пар. В конце приведены выводы и планы дальнейшей работы в данном направлении.

Целью данной работы является анализ существующих мер семантической близости и разработка их модификаций.

1 WordNet

Дж. Миллером и его коллегами из Лаборатории когнитологии Принстонского Университета (США) была разработана модель ментального лексикона человека. Ресурс, который стал первой реализованной глобальной онтологической сетью, по-

лучил название WordNet [6] и со временем стал одним из наиболее авторитетных и распространенных стандартов, используемых для построения лексико-семантических баз.

Популярность и широкое распространение WordNet обусловлены прежде всего его существенными содержательными и структурными характеристиками. Принстонский WordNet и все последующие варианты для других языков направлены на отображение состава и структуры лексической системы языка в целом, а не отдельных тематических областей. Нынешняя версия WordNet охватывает общеупотребительную лексику современного английского языка – более 120 000 слов.

Базовой структурной единицей Принстонского WordNet является синонимический ряд (синсет), объединяющей слова с подобным значением. Каждый синсет представляет в словаре некоторое лексикализированное понятие данного языка. Для удобства использования словаря человеком каждый синсет дополнен дефиницией (gloss) и примерами употребления слов в контексте. Синсеты в WordNet связаны между собой такими семантическими отношениями, как гипонимия (родовидовое), меронимия (часть – целое), лексический вывод (каузация, пресуппозиция) и др.; среди них особую роль играет гипонимия: она позволяет организовывать синсеты в иерархические структуры (деревья таксономии). Лексика каждой части речи представлена в виде набора деревьев (леса). Для разных частей речи родовидовые отношения могут иметь дополнительные характеристики и различаться областью распространения.

Путем между двумя синсетами на WordNet назовем последовательность синсетов, в которой каждая последовательная пара синсетов связана определенным отношением.

2 Меры сходства и связанности

Рассмотренные ниже меры можно условно разделить на основанные на путях (path based) и основанные на описаниях (gloss based). Первые используют кратчайшие пути между концептами в базе знаний, а основанные на описаниях используют словарные описания концептов.

2.1 Основанные на путях

Основанные на путях меры были разработаны в основном только для IS-A отношений – гипо-и гипернимии (конкретизация и абстрагирование). То есть эти меры определены на таксономии. Как отмечается в [7] и у других авторов, большинство таксономий имеют следующий недостаток: один таксономический шаг (таксономическая связь) может быть более мелким, а другой наоборот – более широким. Например, в WordNet между понятиями FORK и SALAD FORK и между FAUNA и CHORDATE одинаковое таксономическое расстояние – одна связь типа IS-A, но интуитивно понятия из первой пары гораздо ближе друг к другу, чем со второй.

PATH

Простейшей основанной на путях мерой является мера, которую будем обозначать *PATH*. Согласно этому подходу, мерой семантической схожести между двумя концептами является обратное значение длины кратчайшего пути в таксономии между этими концептами.

$$sim_{PATH}(c_1, c_2) = \frac{1}{ShortestLength(c_1, c_2)}.$$

LCH

Следующая мера, описание которой приводится здесь, была предложена в [2]. Эту основанную на путях меру семантического сходства, будем обозначать как *LCH*.

При этом подходе мера сходства двух концептов определяется как отношение кратчайшего пути в IS-A иерархии к диаметру таксономии. Для WordNet 2.1 диаметр таксономии существительных равняется 17. Следующая формула описывает меру:

$$sim_{LCH}(c_1, c_2) = -\log \frac{ShortestLength(c_1, c_2)}{2 \times D}.$$

ShortestLength (c_1, c_2) – длина кратчайшего пути (с наименьшим количеством узлов) между концептами c_1 и c_2, а D – это диаметр таксономии.

Авторами была разработана и протестирована модификация этого подхода (далее – *LCH+).* В дополнение к IS-A отношениям в путях допускаются отношения типа «часть-целое» (меронимия и голонимия). Модифицированная мера уже ближе к мере связанности и не является чистой мерой сходства.

WUP

В своей работе по разработке системы машинного перевода английских глаголов на мандаринский китайский [3] предложили следующую меру семантического сходства между концептами c_1 и c_2 :

$$sim_{WUP}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)},$$

где *depth (c)* – это глубина концепта в IS-A иерархии, а *LCS* (c_1, c_2) – ближайший общий родовой узел. Например, в WordNet-таксономии ближайшим общим родовым узлом для узлов «NIKEL» и «DIME» будет «COIN» (рис.1). То есть «NIKEL» и «DIME» объединяет то, что и то, и другое является монетами – «COIN.A» у «COIN» и у «CREDIT CARD» общим является то, что это средства обмена.

2.2 Основанные на описаниях

Подход к определению семантической связанности концептов на основе их словарных описаний был предложен в [4]. Суть данного подхода довольно простая – семантическая связь двух концептов прямопропорциональна количеству слов (или токенов), входящих одновременно в описание первого и второго концепта. Будем обозначать ее LESK. Эта мера является мерой семантической связанности и может быть легко использована для различных частей речи и их комбинаций, в отличие от предыдущих мер, воспринимающих только существительные из-за использования в их определении IS-A иерархии, которая в WordNet разработана лучше для существительных.

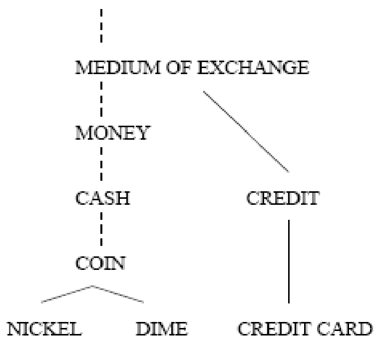


Рисунок 1 – Фрагмент WordNet-таксономии. Сплошные линии представляют IS-A-отношения, а пунктирными показано, что некоторые узлы опущены, чтобы сохранить место

Patwardhan, Banerjee и Pedersen сделали модификацию этого подхода [5]. Будем обозначать его LESK_A (Lesk Adapted). Так как, например, в WordNet описания концептов не очень большие, то их порой не хватает для определения связанности: у двух похожих понятий может не быть ни одного пересечения их описаний. Поэтому эти авторы предложили принимать к рассмотрению не только описание самого концепта, но и непосредственно соединенных с ним концептов в базе знаний.

Так как величина сечения глоссариев считается в словах, то такая мера не является нормализованной и не всегда очевидно показательной (например, длина пересечения в 10 словах может быть лучше, чем длина пересечения в 20, если в первом случае длина обоих глоссариев была, например, по 12 слов, а во втором – по 40). Поэтому применим некоторые соотношения для нормализации. Введем обозначения:

$$I = \left| gloss(c_1) \cap gloss(c_2) \right|$$

$$gloss(c) = \begin{cases} c.gloss, \text{если вычисляем } LESK; \\ \bigcup\limits_{s \in Related(c)} s.gloss, \text{если вычисляем } LESK_A, \end{cases}$$

где *c.gloss* – это описание концепта *c, Related (c)* – это множество непосредственно связанных концептов с концептом *c*. Отметим, что согласно этим обозначениям $sim_{Lesk}^{Simple} = I$ является описанной выше ненормализованной мерой.

Итак, можно привести следующие соотношения:

$$sim_{Lesk}^{Ar}(c_1, c_2) = \frac{\dfrac{I}{\left| gloss(c_1) \right|} + \dfrac{I}{\left| gloss(c_2) \right|}}{2};$$

sim_{Lesk}^{Ar} – это среднее арифметическое соотношений размера пересечения двух описаний к размерам каждого из них. Эта функция принимает значения от 0 до 1.

$$sim_{Lesk}^{Lin}(c_1, c_2) = \frac{2 \times I}{\left| gloss(c_1) \right| + \left| gloss(c_2) \right|};$$

sim^{Lin} – это отношение строится на базе Similarity Theorem [8], фактически, это соотношение количества общей информации о двух объектах к количеству информации о каждом. Принимает значения от 0 до 1.

3 Эксперимент и результаты

Эксперимент был поставлен на двух множествах данных.

Первая – это 353 пары английских слов, не только существительных. Каждой паре, на основе опроса субъективной оценки сходства, были проставлены определенные значения от 0 до 10, где 10 ставился для пары абсолютно одинаковых понятий, а 0 – вообще не похожих и не связанных. Загрузить эти данные можно по адресу http://www.cs.technion.ac.il/ ~ gabr/resources/data/wordsim353/wordsim353.html .

Второе множество – это множество из 30 существительных, все из них есть в WordNet. Эти данные были взяты из статьи [9]. Плюс этого множества данных в том, что на ней было много раз протестировано другими исследователями различные меры, поэтому можно сравнить описанные в этой статье меры и их модификации с другими результатами.

Так как дано пары слов, а методы вычисляют меру концептов, то в этой работе мера между словами вычисляется по формуле:

$$sim_X^{words}(w_1, w_2) = \max_{\substack{c_1 \in w_1 \cdot meanings \\ c_2 \in w_2 \cdot meanings}} (sim_X^{concepts}(c_1, c_2)),$$

где X – название меры, *w.meanings* – это множество смыслов-концептов этого слова.

Для основанных на путях мер проводилось два эксперимента: с максимальной длиной пути в 8 и в 10 узлов.

Для каждой основанной на описаниях меры исчислялись разные соотношения *(Ar, Lin, Simple)*. Кроме того, как было указано выше в разделе описания мер, было реализовано 2 варианта меры LESK_A – с использованием IS-A отношений (обозначим далее LESK_A_ISA) и с использованием IS-A отношений с меронимическими отношениями (обозначим далее LESK_A_MER).

Также была реализована случайная мера (обозначим RAND), которая ставила случайное значение семантического сходства каждой паре. По результатам эксперимента, как видно из таблиц, все меры лучше случайной.

В таблицах приведены значения корреляции различных мер к ответам людей.

Таблица 1 – Результаты эксперимента основанных на путях мер на 353 парах

Максимальная длина пути = 8		Максимальная длина пути = 10	
LCH	0.2654	LCH	0.2654
LCH +	0.2738	LCH +	0.2905
WUP	0.2639	WUP	0.24
PATH	0.3621	PATH	0.3629

Таблица 2 – Результаты эксперимента основанных на описаниях мер на 353 парах

LESK Ar	0.3558	LESK_A_ISA Ar	0.4168	LESK_A_MER Ar	0.4332
LESK Lin	0.3461	LESK_A_ISA Lin	0.4103	LESK_A_MER Lin	0.4199
LESK Simple	0.3233	LESK_A_ISA Simple	0.273	LESK_A_MER Simple	0.2662

Таблица 3 – Результаты эксперимента основанных на путях мер на 30 парах

Максимальная длина пути = 8		Максимальная длина пути = 10	
LCH	0.7927	LCH	0.784
LCH +	0.8125	LCH +	0.8125
WUP	0.6629	WUP	0.6366
PATH	0.7874	PATH	0.7818

Таблица 4 – Результаты эксперимента основанных на описаниях мер на 30 парах

LESK Ar	0.5718	LESK_A_ISA Ar	0.7596	LESK_A_MER Ar	0.766
LESK Lin	0.56	LESK_A_ISA Lin	0.7054	LESK_A_MER Lin	0.7146
LESK Simple	0.4714	LESK_A_ISA Simple	0.4794	LESK_A_MER Simple	0.4019

Таблица 5 – Результаты для случайной меры

На 353 парах		На 30 парах	
RAND	0.030587	RAND	0.227035

В [7] отмечается, что для множества [9] корреляция повторного теста людьми составляет $r = 0.8848$, то есть это можно считать максимально возможной корреляцией.

Выводы

Как мы видим, лучшую корреляцию показала мера ___. Реализованные меры показывают неплохие результаты, достаточные для использования их в других приложениях. Дальнейшие исследования будут сосредоточены на использовании в качестве базы знаний свободной энциклопедии Wikipedia, а также на анализе поведения вышеизложенных мер в других приложениях компьютерной лингвистики.

Литература

1. Budanitsky A. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures / A. Budanitsky, G. Hirst // Workshop on WordNet and other Lexical Resources, Second meeting of the North American Chapter of the Association for Computation Linguistics. – Pittsburg, 2001.
2. Leacock C. Combining local context and WordNet similarity for word sense identification / C. Leacock, M. Chodorow // WordNet: An electronic lexicaldatabase / [ed. C. Fellbaum]. – MIT Press., 1998. – P. 265-283.
3. Wu Z. Verb semantics and lexical selection / Z. Wu, M. Palmer // 32nd Annual Meeting of the Association for Computational Linguistics. – 1994. – P. 133-138.
4. Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone / M. Lesk // Proceedings of SIGDOC. – 1986.
5. Patwardhan S. Using measures of semantic relatedness for word sense disambiguation / S. Patwardhan, S. Banerjee, T. Pedersen // Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. – 2003. – P. 241-257.
6. [Электронный ресурс]. – Режим доступа : http://wordnet.princeton.edu/
7. Resnik P. Using information content to evaluate semantic similarity in a taxonomy / P. Resnik // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – 1994. – P. 448-453.
8. Lin D. An information-theoretic definition of similarity / D. Lin // Proceedings of the International Conference on Machine Learning. – 1998.
9. Miller G. Contextual Correlates of Semantic Similarity / G. Miller, W.G. Charles // Language and Cognitive Processes. – 1991. – Vol. 6, № 1. – P. 1-28.

А.В. Анісімов, К.С. Лиман, О.О. Марченко

Методи обчислення мір семантичної близькості слів природної мови

У даній статті наводяться експериментальні дані обчислення мір семантичної подібності та зв'язаності. Всі міри, що представлені в статті, використовують як джерела знань тільки WordNet. Також авторами були запропоновані й перевірені в експерименті модифікації існуючих мір.

A.V. Ansimov, C.S. Lyman, A.A. Marchenko

The Computational Methods for the Semantic Proximity Measures of Natural Language Words

This article reports about the experimental data on measures of semantic similarity and relatedness computation. All discussed measures use WordNet as a knowledge source. Also, modifications of existing measures were proposed by the authors and were tested in the experiment.

Статья поступила в редакцию 01.07.2010.